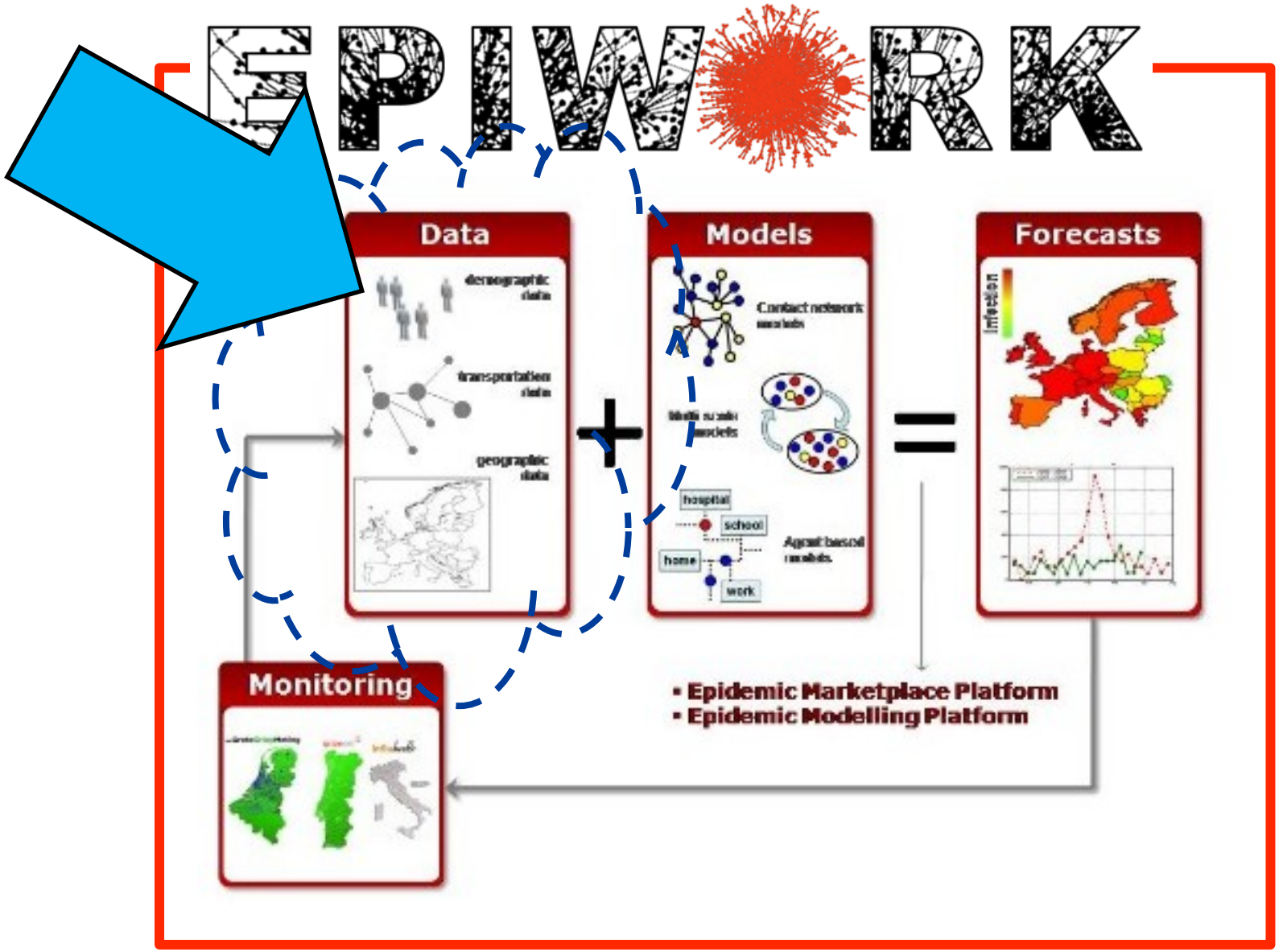
WP3 – Information Platform

Mário J. Silva

Universidade de Lisboa, Faculdade de Ciências,
Departamento de Informática

mjs@di.fc.ul.pt


EPIWORK
Data
demographic data
transportation data
geographic data
Models
Contact network models
Multi scale models
Agent based models
hospital
school
home
work
Forecasts
Infection
Monitoring
• Epidemic Marketplace Platform
• Epidemic Modelling Platform

Data in EPIWORK

- [National Bureau of Statistics] demographics, transportation data, ..
- [Public Health authorities] surveillance data (maybe?)
- **[Internet Social Networks] behavioural data**

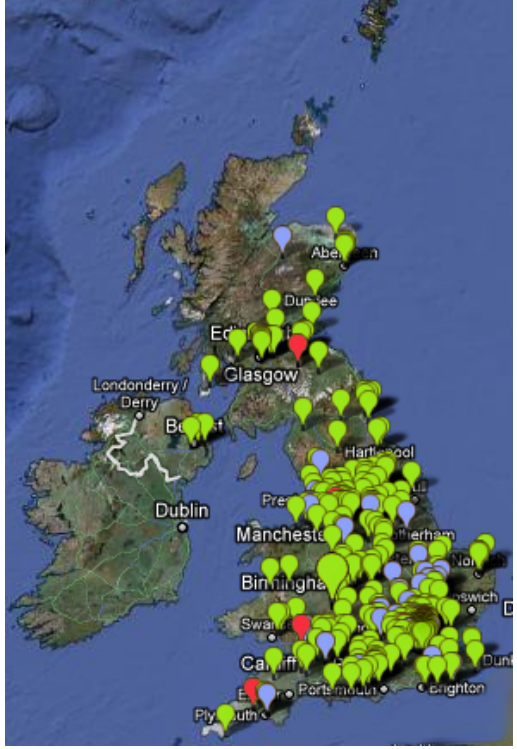
To be shared by epidemic modellers in a digital library, dubbed the **Epidemic Marketplace**

What will be necessary to predict epidemics precisely?

- **Data** of many different types and many unrelated sources.
  - Improved accuracy makes required data a never-ending story
  - We all want to see realistic and timely plots of epidemics propagation.
  - Available, but **hard to find, collect and maintain**!

Número de participantes

Portugal	5477 participantes
Holanda	20906 participantes
Bélgica	5839 participantes
Itália	3019 participantes
México	3950 participantes
Brasil	370 participantes
Reino Unido	5551 participantes
Australia	7776 participantes
Montreal (CA)	951 participantes



gripenet

http://www.gripenet.pt/

EPIWORK

INFLUWEB

Other Internet Monitoring Sources

disease	location	source	author	evidence	score	date
HIV	Italy	Twitter	HumanityNews (Humanitarian News)	News: Patent pool decision heralds era of cheap **HIV** drugs http://dlvr.it/L5Bq	0	2010-03-22 10:48:14
HIV	Italy	Twitter	USArmyAfrica (USArmyAfrica)	News Update Africa: Steps in the right direction against HIV/**AIDS**: Africa's battle against HIV/**AIDS** is not ... http://bit.ly/acEbGC #Africa	0	2010-03-22 09:19:32
HIV	Italy	Twitter	romavisibile (Roma Visibile)	Social Romavisibile L'**Aids** lavorando per il guru e la setta mi ha abbandonata http://ow.ly/16Rhcl	0	2010-03-22 09:16:44
HIV	Italy	Twitter	Grantun19 (Grant Vaughn)	Safety concern for **HIV** drug combination http://is.gd/aSBEL	0	2010-03-22 08:37:05
HIV	Italy	Twitter	NonProfitBlogs (NonProfitBlogs)	**Aids** warning over bushmeat trade http://dlvr.it/L31N	0	2010-03-22 06:54:17
HIV	Italy	Twitter	HumanityNews (Humanitarian News)	Blogpost: **Aids** warning over bushmeat trade http://dlvr.it/L2Gb	0	2010-03-22 05:22:13
HIV	Italy	Twitter	casso89 (Antonino Cassotta)	Farmaco contro l'acne previene riattivazione di **HIV** latente - http://wp.me/pObs6-6l	0	2010-03-21 22:01:25
HIV	Italy	Twitter	mwbloem (Martin Bloem)	GWU and WFP organized a meeting on nutrition/food security in HIV/**AIDS** programs; excellent input to the new WFP's policy on HIV **AIDS**.	0	2010-03-21 19:08:38
HIV	Italy	Twitter	duduramone (EDUARDO N. VANTINI)	@niderodriguez tomara que vc morra corroido pela **aids**	0	2010-03-21

Other Internet Monitoring Sources

google.org Flu Trends

Google.org home

Flu Trends

Japan

Download data

Home

How does this work?

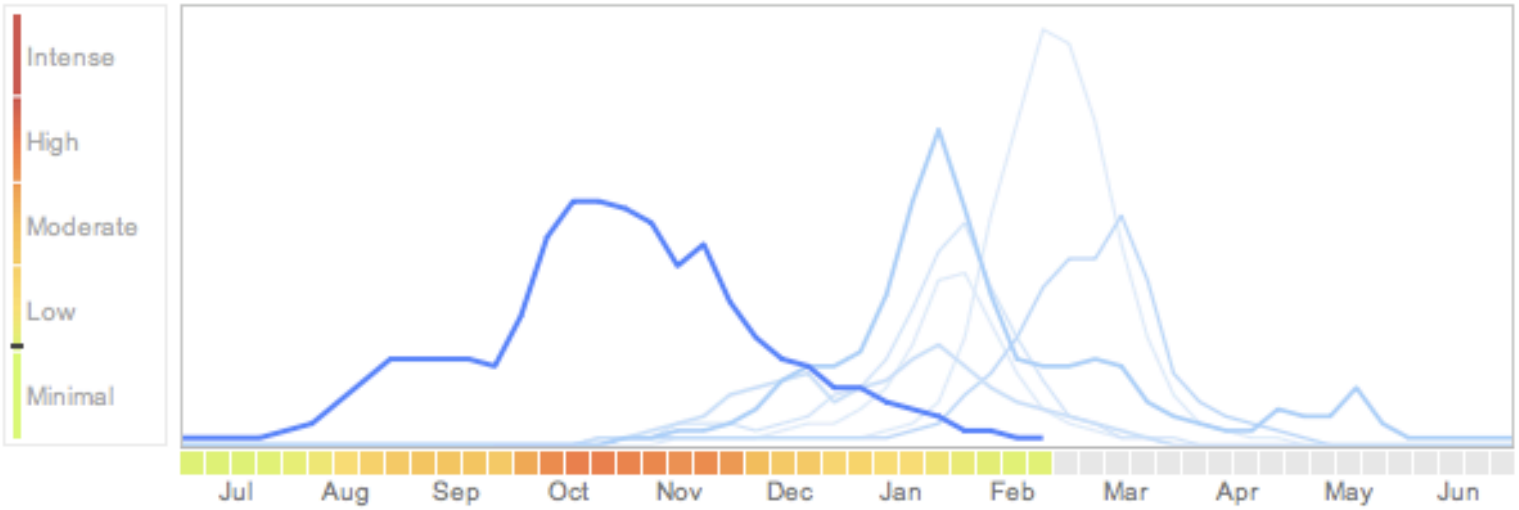
FAQ

Explore flu trends - Japan

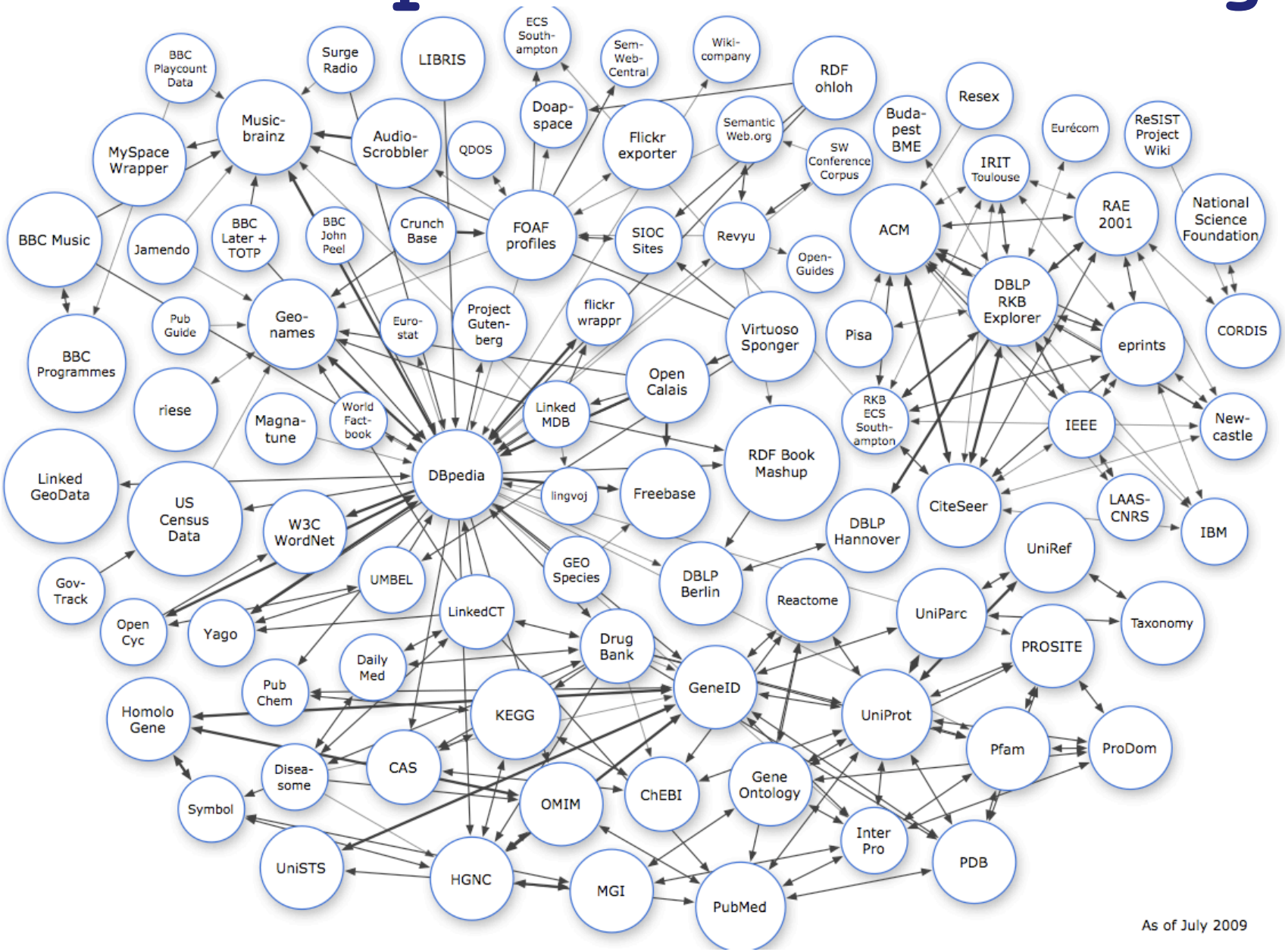
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

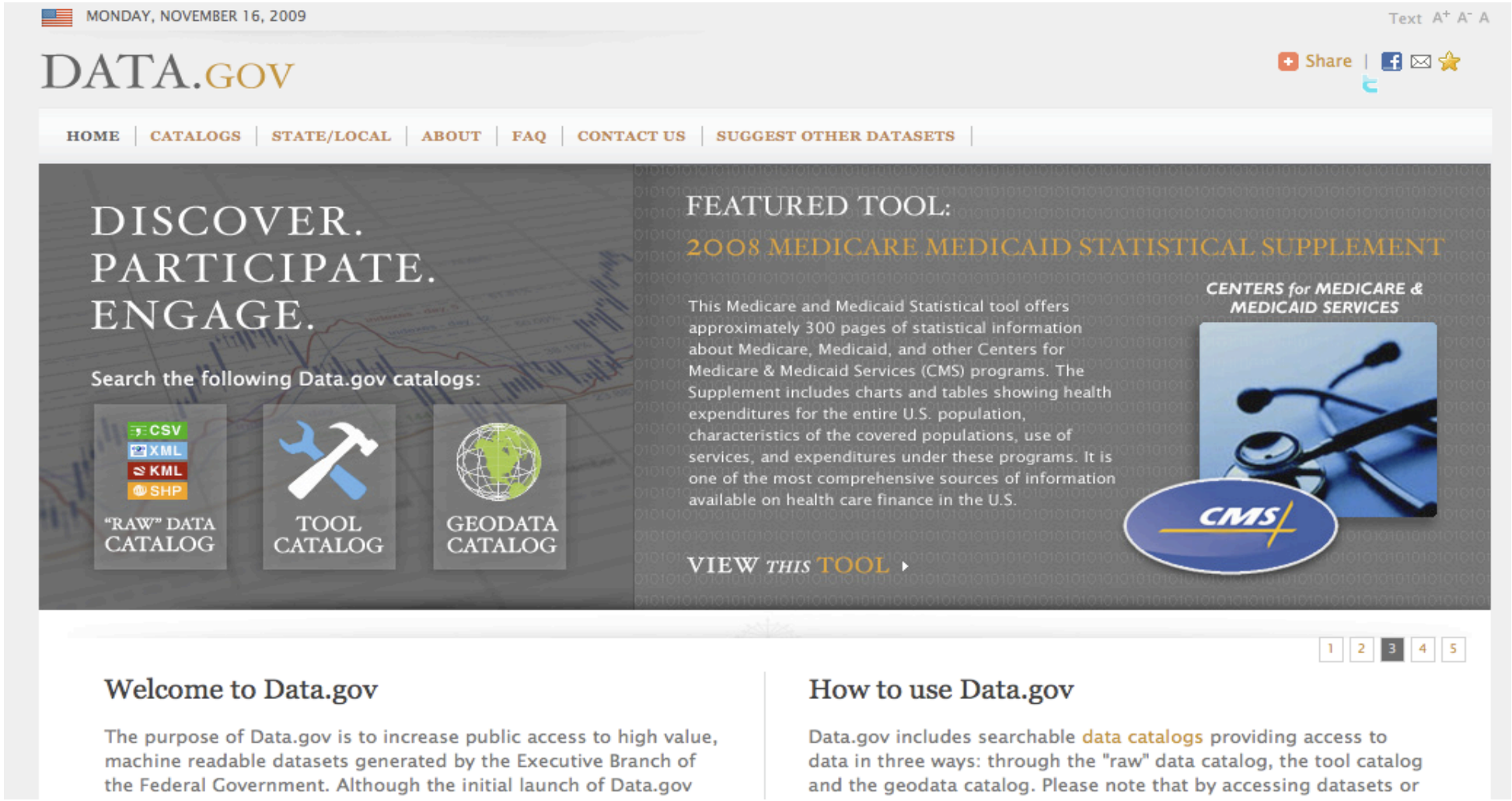
2009-2010 Past years



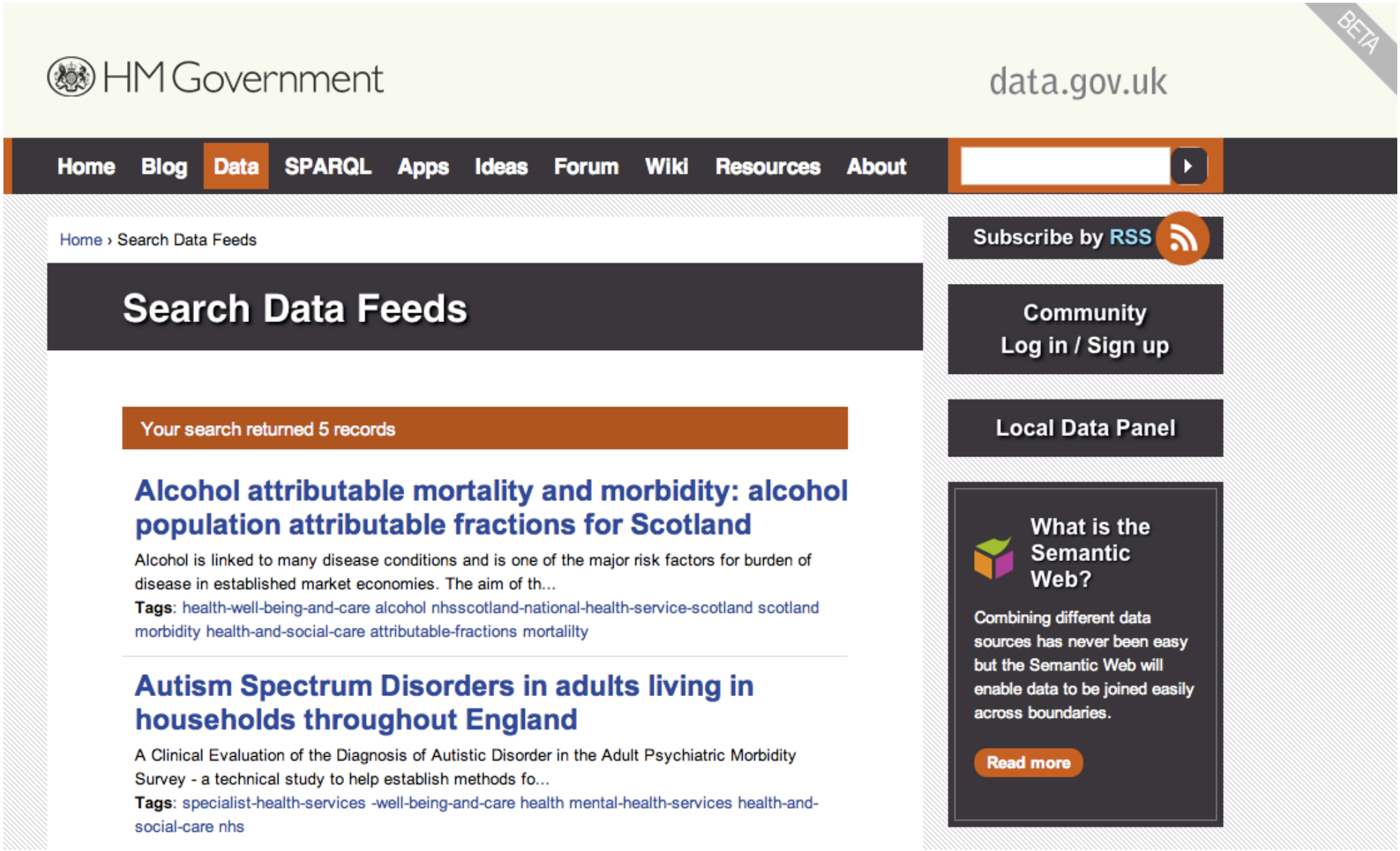
http://linkeddata.org/



As of July 2009


MONDAY, NOVEMBER 16, 2009
Text A+ A- A
DATA.GOV
Share
HOME | CATALOGS | STATE/LOCAL | ABOUT | FAQ | CONTACT US | SUGGEST OTHER DATASETS
DISCOVER.
PARTICIPATE.
ENGAGE.
Search the following Data.gov catalogs:
CSV
XML
KML
SHP
"RAW" DATA CATALOG
TOOL CATALOG
GEODATA CATALOG
FEATURED TOOL:
2008 MEDICARE MEDICAID STATISTICAL SUPPLEMENT
This Medicare and Medicaid Statistical tool offers approximately 300 pages of statistical information about Medicare, Medicaid, and other Centers for Medicare & Medicaid Services (CMS) programs. The Supplement includes charts and tables showing health expenditures for the entire U.S. population, characteristics of the covered populations, use of services, and expenditures under these programs. It is one of the most comprehensive sources of information available on health care finance in the U.S.
VIEW THIS TOOL
CENTERS for MEDICARE & MEDICAID SERVICES
CMS
1 2 3 4 5
Welcome to Data.gov
The purpose of Data.gov is to increase public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government. Although the initial launch of Data.gov
How to use Data.gov
Data.gov includes searchable data catalogs providing access to data in three ways: through the "raw" data catalog, the tool catalog and the geodata catalog. Please note that by accessing datasets or

http://data.gov.uk/data/list?keyword=epidemiology



Epidemic Marketplace (EM)

1. **Catalogue of data sources** containing the metadata describing existing databases;
2. **Forum** to
   - publish information about data
   - seek modellers to collaborate with,
   - seek sources of data that could be of interest to their epidemiological modelling efforts;
3. **Mediating software** to automatically process queries to epidemiological data, harvest data, assemble datasets....

Outline

1. The need for an Epidemic Marketplace
2. **Metadata and Ontologies for Epidemic Modelling (Deliverable D3.1)**
3. Epidemic Marketplace Architecture & Implementation (Deliverable D3.2)
4. Where we stand and forecasts

Steps for Creating the EM

1. Elaborate **meta-model** for describing datasets used by epidemic modellers.
2. Provide **query services** over the meta-data to discover resources.
3. Select **ontologies** for characterizing data and develop an ontology of epidemic concepts.
4. Ingest, harmonize and **cross-link data.**
5. Provide **query services to select epidemic data** using the EM meta-data and ontologies.

Common Reference Model

- **Open domain:** detailed description of the datasets used in the models of all sorts of epidemics would **require describing virtually every kind of information**, given the diversity of factors and the interdisciplinary of epidemiologic studies.

Data model needs to support **interlinked data.**

Meta-data and Ontologies

- The **information model** of the EM is directly **defined as metadata and ontologies**.
- Advantages of using a specific ontology to describe a specific disease
  - makes everybody referring to a specific disease to use the same term, making the **information discovery simpler** and more complete;
  - keeps the **metadata text simpler**, the ontology itself contains other data that doesn't need to be inserted as metadata

Metadata standards

- ISO/IEC 11179 Metadata Registry (MDR)
- **Dublin Core (DC)**
  metada for the Web, 15 properties
  - ISO Standard Standard 15836-2003 of February 2003, ANSI/NISO Standard Z39.85-2007 of May 2007 and IETF RFC 5013 of August 2007.
  - **DCMI namespace**: Since 2008, DCMI includes formal domains and ranges in the definitions of its properties.

Strategies for Creating an Epidemic Data Metadata Model

- Start with a catalogue of epidemic datasets…
- Focus on collecting extensive metadata.
- Leverage ontologies and their technologies
  - establish the common terminology
  - interlink heterogeneous metadata classifications.
  - **connect with the OBO (Open Biomedical Ontologies) initiative**

Strategies for Creating a Metadata Model for Epidemic Data (II)

- **Ontologies** will serve to integrate heterogeneous data sources as they provide **semantic relationships among the described objects**.
- Further on, the **EM will include methods and services for aligning the ontologies**.
- We expect that this can spawn a **virtuous cycle**, stimulating the cataloguing and linking by the epidemic modellers community.

Strategies for Creating a Metadata Model for Epidemic Data (III)

- With DCMI terms and conventions + Linked Data conventions**, turn datasets into web resources.**
  - **describe the data** structures in the datasets **using ontologies**.
  - descriptions will be used by **people** and **information discovery tools**

Strategies for Creating a Metadata Model for Epidemic Data (IV)

- Define policies establishing the **level of detail of the metadata.**
  - low level of detail may not be able to sufficiently describe the datasets, making the right information harder to find
  - a too detailed metadata scheme can turn the annotation of a specific dataset into a daunting task, hindering the acceptance of the model by the user community.

Strategies for Creating a Metadata Model for Epidemic Data (V)

- Started modelling the datasets with low detail, annotating the 15 standard DC elements as **character data.**
- Further down the line, we initiate the annotation of DC elements with **semantically richer descriptions**
- Metadata **annotation criteria have to follow a common standard**, so data can be comparable and searched using similar queries
  - **use controlled languages as much as possible** and languages for describing data structures, progressively limiting the use of free text.

Strategies for Creating a Metadata Model for Epidemic Data (VI)

Analysed selected

sample of datasets

- EM Twitter Datasets: harvested with software prototype of the EM.
- US Airports Dataset: Data about the airport network of the United States.

Surveyed published articles in epidemiology journals and inferred the attributes of the used datasets

- We annotated datasets to which we did not actually have access, but devised what would be their metadata description as DC elements.

Outline

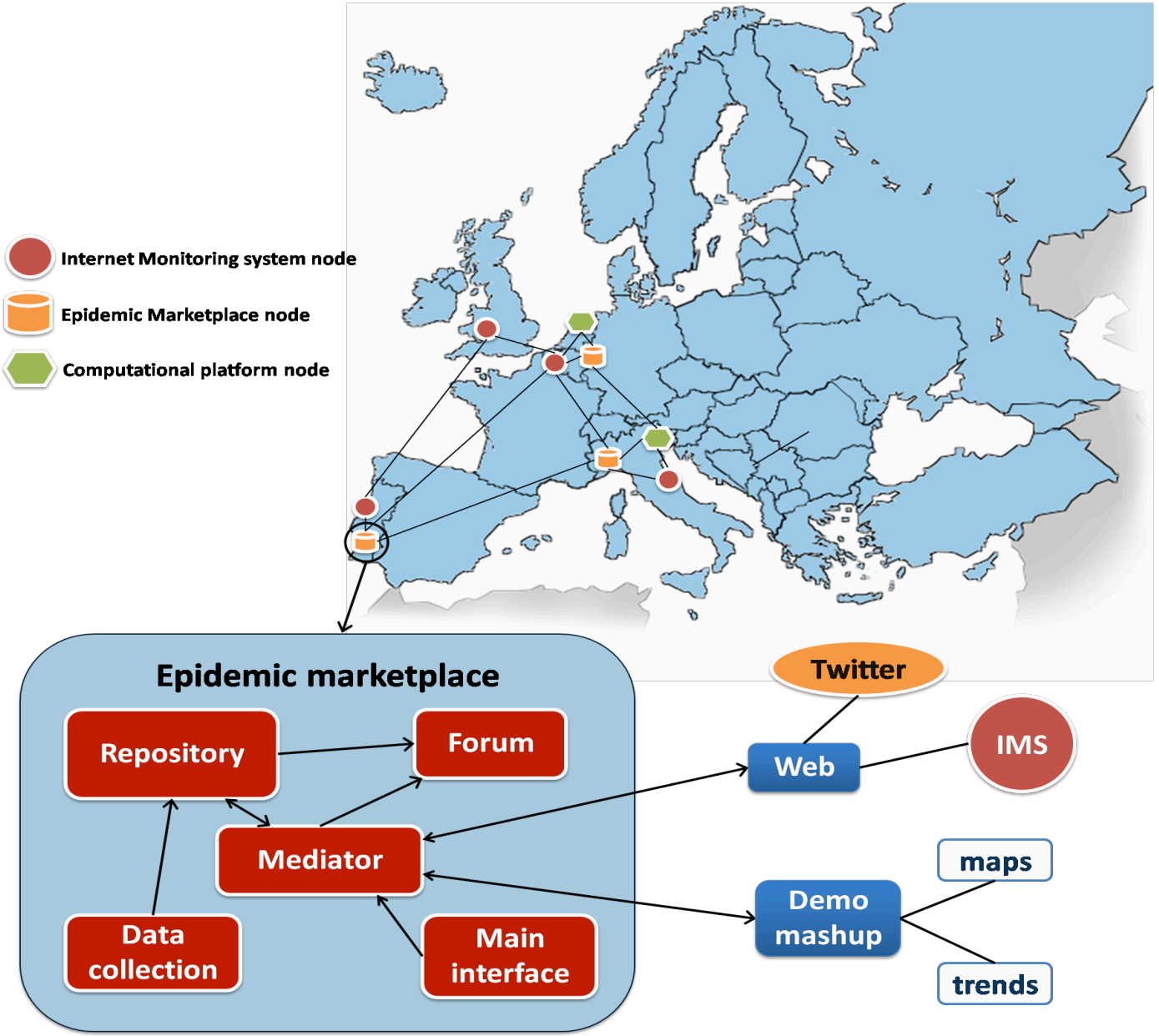
1. The need for an Epidemic Marketplace
2. Metadata and Ontologies for Epidemic Modelling (Deliverable D3.1)
3. **Epidemic Marketplace Architecture & Implementation (Deliverable D3.2)**
4. Where we stand and forecasts

The EM as a Virtual Repository

- The Epidemic Marketplace is composed of a set of, geographically distributed, **interconnected data management nodes**, sharing**:**
  - **common data models**,
  - an **authorization infrastructure**
  - **access interfaces**.
- At each node, a set of software components implements a set of requirements that characterize their performance and interfaces.

EM: Main Components

- **Repository:** Stores epidemic data sets and ontologies to characterise the semantic information of the data sets.
- **Mediator:** A collection of web services that will provide access to internal data and external sources, using state-of-the-art semantic-web/grid technologies.
- **Collector:** Retrieves information of real-time disease incidences from publicly available data sources, such as social networks;.
- **Forum:** Allows users to organize discussions centred on the datasets fostering collaboration among modellers.


Internet Monitoring system node
Epidemic Marketplace node
Computational platform node
Epidemic marketplace
Repository
Forum
Mediator
Data collection
Main interface
Twitter
Web
IMS
Demo mashup
maps
trends

EM: Main System Requirements

EM needs to define policies and provide services for:

- Sharing and management of epidemiological data sets.
- Seamless integration of heterogeneous data sources.
- Creation of a virtual community for epidemic research.
- Distributed Architecture.
- Secure access to data.
- Support for data analysis and simulation in grid environments:.
- Workflows

EM Repository Requirements

- **Separation of data and metadata**
  - metadata may contain information not directly accessible.
- **Support for Metadata standards**
  - Dublin Core, because that's what everyone seems to be using
- **Ontology support**
  - for describing and characterising the data.

EM Mediator Requirements

Responsible for data exchanges with **Clients, IMS** and other **data providers** (RSS ProMed Mail, ..):

- **Query and search capabilities on heterogeneous datasets:** in epidemic modelling, diversity is unlimited.
- Access to “plug-in-able” resources:.
- **RESTful interfaces**.

Collector Requirements

- **Active data harvesting:** focused web crawler, subscription of newsfeeds and email services.
- **Passive data collection:** EM preserves and distributes deposited datasets originating from IMS
- **Local storage capability:** all collected data in at least one EM site.
- **Meaningful data partitioning policies:** to epidemic modellers and accounting for legal/administrative barriers

Outline

1. The need for an Epidemic Marketplace
2. Metadata and Ontologies for Epidemic Modelling (Deliverable D3.1)
3. Epidemic Marketplace Architecture & **Implementation** (Deliverable D3.2)
4. Where we stand and forecasts

Software Components

- Fedora Commons for the implementation of the main features of the repository.
- Access control in the platform
  - XACML (OASIS 2010),
  - LDAP (Tuttle et al. 2004)
  - Shibolleth (identity management).
- Front-end based in Muradora
  - now being replaced by the **Drupal** CMS.

Contact

Welcome fedoraAdmin | Your Account | Logout

EPIWORK

Home | Browse | Search | Submit | Publish | Portfolio | Admin Tools

Home

Resource Metadata

Meta-data	
Title	Twitter dataset H1N1 + Portugal 4-6-2009
Creator	LASIGE node of the Epidemic Marketplace
Subject	twitter message dataset
Description	This dataset contains Twitter messages containing the words H1N1 and Portugal collected between 16-5-2009 and 3-6-2009, Information is a 7 columns relation, containing the following data: Column 1- keyword 1 (disease)- H1N1 Column 2- Keyword 2 (location)- Portugal Column 3- Source (Twitter) Column 4- Author of the message (user id) Column 5- The message body (evidence) Column 6- score Column 7- date (day and hour)
Publisher	Epiwork – http://www.epiwork.eu
Format	text/tab-separated-values
Language	English, Portuguese
Contributor	Luis F Lopes, Joao M Zamite, Bruno C Tavares, Francisco M Couto, Fabricio Silva, Mario J Silva
Relation	Luis F. Lopes, João M. Zamite, Bruno C. Tavares, Francisco M. Couto, Fabrício Silva and Mário J. Silva. (2009). Automated Social Network Epidemic Data Collector. INForum informatics symposium.
Source	http://epiwork.di.fc.ul.pt/collector/
Coverage	Spatial: Portugal, Temporal: 2009-5-16 to 2009-6-3
Rights	Creative Commons Attribution-ShareAlike (CC BY-SA), http://creativecommons.org/licenses/by-sa/3.0/

Current Focus

- Refining and populating**, *enriching* the catalogue** of epidemic resources **using initial prototype.**
  - The method of scanning published epidemic modelling studies and then inferring the metadata descriptions has shown to be very useful.
- **Designing the user interface for the second version**.
  - Must be useful to the expert and occasional user.

Forthcoming Developments

- Identifying ontologies (and ontology terms) to use. Linking to ontology definition initiatives.
- Linking ontologies and web data **using linked data** conventions and **ontology alignment** methods.

Outline

1. The need for an Epidemic Marketplace
2. Metadata and Ontologies for Epidemic Modelling (Deliverable D3.1)
3. Epidemic Marketplace Architecture & Implementation (Deliverable D3.2)
4. **Where we stand and forecasts**

WP3: status

- Deliverable D3.1 (meta-model) released
- Deliverable D3.2 (prototype) released
  - Hardware and base software deployed;
  - Initial prototype of EM with initial set of characterized datasets
- Overcoming the initial difficulties in hiring the planned resources.

Publications in the 1st year

1. Mário J. Silva, Fabrício A.B. Silva, Luís Filipe Lopes, Francisco M. Couto, **Building a Digital Library for Epidemic Modelling**. Proceedings of ICDL 2010 - The International Conference on Digital Libraries 1, p. 447--459, New Delhi, India, 23--27 February, 2010. TERI Press -- New Delhi, India. Invited Paper.
2. Luis Filipe Lopes, João Zamite, Bruno Tavares, Francisco Couto, Fabrício A.B. Silva, Mário J. Silva, **Automated Social Network Epidemic Data Collector**. INForum - Simpósio de Informática September, 2009.

Current Challenges

- Motivate the community to populate the Epidemic Marketplace.
  - Chicken and egg situation.
- **Data anonymization** is a major concern
  - Rights management to the sentence level!
  - Anyone giving away curated UGC?
- Access control policies
- Dataset selection and generation policies

New Poll: Is anonymization of large datasets still possible?

Although the first Netflix prize was very successful, Netflix recently **cancelled the 2nd Netflix Prize** after FTC expressed privacy concerns and a lawsuit was filed.

The privacy concerned were sparked by 2 researchers, **Arvind Narayanan and Vitaly Shmatikov**, who showed that some people who rank obscure movies can be identified by correlating their rankings to other online information, such as IMDB rankings

Although these researchers wrote an **open letter to Netflix** arguing for more research and privacy-preserving data analysis mechanism, it seems doubtful that this would happen.

The new KDnuggets Poll is asking:

Is it still possible for companies like Netflix to anonymize large datasets?

Please vote on **www.kdnuggets.com**

See also **How Privacy Vanishes Online** (New York Times), which makes the point

Latest News

- **New Poll: Is anonymization still possible?**
- **Eventbrite: Analytics Engineer**
- **Google Analytics to allow opt out**

NEW **KDnuggets 10:n06: Is Privacy still possible online? KDD Cup; KDD Workshops**

WP3 SWOT Analysis

Strengths

- Epiwork-driven EM
- Standards-based
- Open Source modules
- Supported (until 2012)

Weaknesses

- **Unpopulated EM**
- Looking for the right policies
- What are the incentives?
- Interfaces to WP4 and WP5?

WP3 SWOT Analysis

Opportunities

- Epiwork testbed
- Creation of a baseline for epidemic modelling
- Showcase for partners' outputs

Threats

- Consortium enters "everyone for himself" mode.
- "Somebody will take care of that" attitude
- EM perceived as a very expensive, complex and useless cache

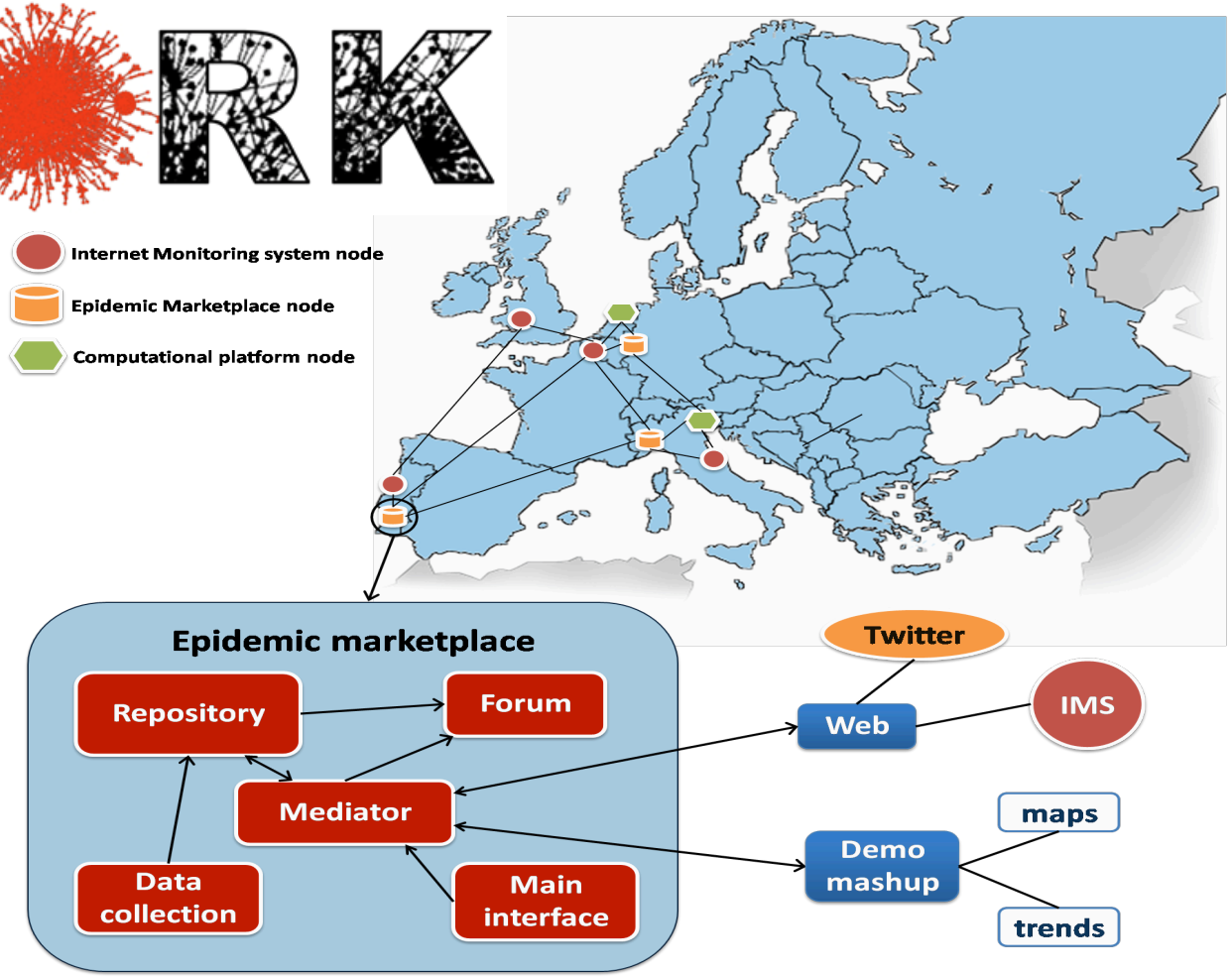
Todo list and planning

1. Populate Repository
2. Linked Epidemic Data
3. Ethics, Privacy and Anonimization
4. Access control policies
5. Dataset selection generation
6. Distributed Authentication
7. Replicate EM node

Scheduled Deliverables

		Year 1			Year 2			Year 3			Year 4		
		M4	M8	M12	M16	M20	M24	M28	M32	M36	M40	M44	M48
WP3	**Information platform**												
Task 1			**D3.1**										
Task 2			**D3.1**										
Task 3				**D3.2**		**D3.3**				**D3.4**			**D3.6**
Task 4										**D3.4** **D3.5**			**D3.6**

EPIWORK



SEVENTH FRAMEWORK PROGRAMME

http://www.epiwork.eu